





RESEARCH ARTICLE 

# Sense-aware connective-based indices of cohesion and their relationship to cohesion ratings of English language learners' written production

Xiaofei Lu<sup>1</sup>  and Renfen Hu<sup>2</sup> 

<sup>1</sup>The Pennsylvania State University; <sup>2</sup>Beijing Normal University  
**Corresponding author:** Renfen Hu; Email: [irishu@mail.bnu.edu.cn](mailto:irishu@mail.bnu.edu.cn)

(Received 10 June 2023; Revised 10 February 2024; Accepted 21 February 2024)

## Abstract

The use of connectives has been considered important for assessing the cohesion of written texts (Crossley et al., 2019). However, existing connective-based indices have not systematically addressed two issues of ambiguity, namely, that between discourse and non-discourse use of polysemous word forms and that in terms of the specific discourse relations marked by polysemous discourse connectives (Pitler & Nenkova, 2009). This study proposes 34 sense-aware connective-based indices of cohesion that account for these issues and assesses their predictive power for cohesion ratings in comparison to 25 existing indices. Results from the analysis of 3,911 argumentative essays from the English Language Learner Insight, Proficiency and Skills Evaluation Corpus show that 23 sense-aware indices but only three existing indices correlated significantly and meaningfully with cohesion ratings. The sense-aware indices also exhibited greater predictive power for cohesion ratings than existing indices. The implications of our findings for future cohesion research are discussed.

## Introduction

Recent writing and assessment research has increasingly attended to the role of cohesion in the quality of second language (L2) written production and the ways in which cohesion can be measured in valid and reliable ways (Crossley et al., 2016b, 2019; Zhang et al., 2022). Cohesion is generally understood as an objective property of the explicit text that compasses the linguistic features and devices used to connect different ideas in and parts of a text; it differs from but is closely related to coherence, which refers to the overall level of connectedness, including logic, unity, and comprehensibility, of a text that is evident to its readers (Graesser et al., 2004; Halliday & Hasan, 1976). As detailed in the next section, there is good research consensus that cohesion can be assessed at three levels, namely, local cohesion, global cohesion, and text cohesion, each of which can be measured using one or more different types of indices, such as connectives, lexical overlap, semantic overlap/similarity, givenness, type-token ratio

(TTR), and lexical density (Crossley et al., 2016b, 2019). A sizable body of L2 writing and assessment studies have shown that these different levels of cohesion and/or different types of cohesion indices can significantly predict cohesion, coherence, or quality ratings of L2 written production, although their predictive power may be affected by different learner-related (e.g., proficiency level) and task-related variables (e.g., genre) (Guo et al., 2013; Zhang et al., 2022).

Despite the progress made in conceptualizing and assessing cohesion, a notable gap in existing cohesion research lies in the lack of systematic attention to lexical ambiguity in word-based indices, such as those based on connectives and lexical/semantic overlap. In the case of connective-based indices, the concern of the current study, two issues related to ambiguity exist. The first has to do with the ambiguity between discourse and non-discourse use of polysemous word forms. For example, the word *once* may be used as a discourse connective expressing a temporal relation (e.g., *I will leave once I am done*) or as an adverb meaning “one time” (e.g., *The bell will ring once*) or “previously” (e.g., *I once really liked it*). Recent research has started to attend to but has not yet fully addressed this issue (Crossley et al., 2019) (see discussion in the next section). The second has to do with the specific discourse relation senses that polysemous discourse connectives may be used to express in context. For example, as a discourse connective, the word *since* may be used to express either a temporal relation (e.g., *I haven't seen him since we met in May*) or contingency (*You don't have to go since you are so busy*). This issue has not yet been explicitly or systematically considered in existing connective-based cohesion indices. It would appear reasonable to argue that systematic resolution of these two issues of ambiguity will yield a more accurate characterization of the use of connectives as cohesive devices in written texts and greater reliability of connective-based indices of local cohesion. Considering this research gap, the current study proposes a comprehensive set of 34 sense-aware connective-based cohesion indices that distinguish discourse and non-discourse connective forms as well as four discourse relation senses of discourse connectives (i.e., elaboration, expansion, contingency, and temporal). The study further evaluates the extent to which the proposed sense-aware indices correlate with and predict human cohesion ratings of written texts produced by young English Language Learners (ELLs) in comparison to 25 existing connective-based cohesion indices.

### Measuring cohesion of written texts

Text connectedness (i.e., the degree to which different components of a text, such as clauses, sentences, paragraphs, and sections, and the information contained therein are linked) plays a critical role in the comprehensibility and processing of a text (Halliday & Hasan, 1976). Central to text connectedness are the concepts of cohesion and coherence. Cohesion refers to the use of explicit linguistic and textual features to connect the ideas in different parts of a text, whereas coherence is understood not as an explicit or objective textual property but as the extent to which the text allows its readers to construct a coherent, connected mental representation of its content (Halliday & Hasan, 1976). Different from cohesion, coherence cannot be achieved by linguistic or textual features alone but may interact with variables beyond the text itself, such as the reader's background knowledge, reading skills, and language proficiency (e.g., McNamara et al., 2014).

Scholars have proposed several frameworks or taxonomies of textual cohesion. Halliday and Hasan's (1976) highly influential framework of textual cohesion presents

five types of cohesive ties, “the means whereby elements that are structurally unrelated to one another are linked together, through the dependence of one upon another for its interpretation” (p. 27). The first four types realize what they call grammatical cohesion. These include reference (i.e., the use of personal pronouns, demonstratives, and comparatives to refer back to previously mentioned entities), substitution (i.e., the use of words such as *do* to replace a previous expression), ellipsis (i.e., the omission of expressions implied by the context), and conjunction (i.e., the additive, adversative, causal, and temporal conjunctive relations between sentences signaled by conjunctive elements). The last type realizes lexical cohesion, manifested in reiteration, the use of words with certain lexical relations (e.g., synonyms, hyponyms, and hypernyms), and collocational items (i.e., words that frequently co-occur).

Although echoing the distinctions made between grammatically and lexically driven cohesion and among different types of conjunctive relations by Halliday and Hasan (1976), Louwse (2002) proposed the additional view that cohesion can be achieved at local, global, and text levels. Local cohesive indices are used to connect clauses or sentences, such as explicit connectives (e.g., *while*, *therefore*), lexical overlap between sentences, and semantic overlap/similarity between sentences. Global cohesive indices are used to connect paragraphs in a text, such as lexical overlap between paragraphs and semantic overlap/similarity between paragraphs. Text cohesive devices are used to build connections throughout the text, such as *givenness* features (e.g., the use of a pronoun to refer to a noun referent after its initial mention). With the high level of operationalizability of this view, numerous indices that tap into the use of different types of cohesive devices have been proposed to measure cohesion at these levels, such as those integrated in two widely used computational tools for cohesion analysis, namely, Coh-Metrix (Graesser et al., 2004) and the Tool for the Automatic Analysis of Cohesion (TAACO) (Crossley et al., 2016b, 2019). For example, TAACO 2.0 includes 25 connective-based indices (e.g., number of causal connectives), 108 indices of lexical overlap between sentences (e.g., number of lemma types that occur at least once in the next sentence) or paragraphs (e.g., number of lemma types that occur at least once in the next paragraph), 16 indices of semantic overlap/similarity between sentences or paragraphs (e.g., average sentence to sentence overlap of noun synonyms), four *givenness* indices (e.g., number of third-person pronouns divided by number of nouns), and 15 TTR (e.g., lemma TTR) and lexical density indices (e.g., ratio of content words). It also includes 26 source text similarity indices (e.g., percentage of unigrams in the text that are keywords) for evaluating the similarity between a source text and a target text in source-based writing tasks.

A substantial body of second language acquisition (SLA) research has argued for and offered empirical evidence of the role of textual cohesion in L2 comprehension and production, two critical aspects of SLA. Indeed, cohesive devices can help L2 learners establish connections between ideas and follow the information flow more easily in understanding a L2 text. They also serve as a necessary tool for L2 learners to express their thoughts logically and coherently in L2 production. SLA research on the role of textual cohesion in L2 comprehension has reported that L2 readers rely on referential and lexical cohesion to a far larger extent than first language (L1) speakers in comprehension processes (Jonz, 1987), that L2 readers benefit from causal markers (Degand & Sanders, 2002) and awareness of lexical cohesive links in reading comprehension (Bayraktar, 2011), and that content word overlap affects L2 reading not only in localized processing but also in overall comprehension (Biler, 2018). On the other hand, L2 learners have been found to exhibit more homogeneous processes of meaning representation when reading high-cohesive texts and more heterogeneous ones when reading

low-cohesive texts (Bilki, 2014). These findings confirm that different types and density levels of cohesive devices can affect L2 learners' comprehension and meaning representation processes in multifaceted ways.

Just as L2 learners' comprehension processes are affected by different types and density levels of cohesive devices, the types and density levels of cohesive devices used by L2 learners could be expected to affect the coherence and comprehensibility of their L2 production, which could in turn affect ratings of the cohesion, coherence, or overall quality of their production in the context of L2 assessment. A number of L2 writing studies have reported findings precisely on the relationship of different types of local, global, and text cohesion indices to human ratings of cohesion, coherence, or writing quality. These findings have shown that different levels of cohesion may be related to cohesion, coherence, or quality ratings in different ways. For example, several studies have reported that local cohesion indices (e.g., conditional connectives and lexical overlap) tend to be negatively correlated with quality ratings of L2 writing, whereas global cohesion indices (e.g., lexical overlap between paragraphs) tend to show positive correlations (e.g., Crossley & McNamara, 2012; Guo et al., 2013; Kim & Crossley, 2018). These findings point to the need to disentangle the effect of different levels of cohesion on writing quality. However, some studies have found variation in how different types of indices at the same level of cohesion may be related to cohesion, coherence, or quality ratings. For example, Crossley et al. (2016a) reported that while some connective-based local cohesion indices (e.g., positive intentional connectives) showed negative correlations with quality ratings of timed descriptive essays written by college-level L2 English learners, several others (e.g., positive causal connectives) showed positive correlations, highlighting the importance to differentiate among different types of connectives in using connective-based cohesion indices. Genre appears to be another important factor that needs to be considered in examining the relationship between indices of cohesion and cohesion, coherence, or quality ratings. For example, Zhang et al. (2022) examined the relationship of six cohesion indices from TAACO to the quality ratings of two genres of writing by L1 Chinese college-level learners of English, namely, application letters and argumentative essays. They reported positive correlations for two text cohesion indices (moving-average TTR and the lexical density of word types) for both genres, negative correlations for lemma overlap between adjacent sentences and paragraphs for argumentative essays, and no significant correlation for either positive or negative logical connectives for either genre. Taken together, these studies have showcased the increasing attention to the validation of cohesion indices and the relationship of such indices to human ratings of cohesion, coherence, or quality in recent L2 writing research. They also informed our focus on a specific type of indices at one level of cohesion (i.e., connective-based indices of local cohesion) in the current study, our attention to different types of connectives, and the use of essays of a single genre (i.e., argumentative essays) in our analysis.

Notably, existing cohesion research has not yet systematically attended to issues of lexical ambiguity in developing and validating indices based on word forms. For example, indices based on lexical or semantic overlap rely on matches of either the same lemma forms (in lexical overlap indices) or lemma forms of synonyms (in semantic overlap indices). However, even within the same text, the same word may be used with different meanings, and the degree of lexical ambiguity within the text may affect the reliability of such lexical and semantic overlap indices. As a step toward addressing this research gap, the current study focuses on resolving two issues of ambiguity for connective-based indices, a type of index that has been used extensively in cohesion research. As mentioned previously, the first issue involves the ambiguity

between discourse and non-discourse use of polysemous word forms (e.g., *so* as a connective and an intensifier). Although prior cohesion research largely ignored this issue, Crossley et al. (2019, p. 17) used the Stanford Neural Network Dependency Parser (Chen & Manning, 2014) “to disambiguate word forms that can be used as both cohesive devices and for other purposes” in developing TAACO 2.0. However, as specified in the TAACO 2.0 manual found at the TAACO website,<sup>1</sup> in each of the various lists of different types of connectives, only a small number of connectives that receive the “mark” tag in the Stanford dependency representation are disambiguated. For example, the list of “positive causal connectives” contains 41 items, among which only two items (*since*, *so*) are disambiguated this way, whereas many other potentially ambiguous word forms (e.g., *condition*, *due*, *even*, *follow*, *make*, *only*, *following*) are not. As such, the effort to resolve the first ambiguity issue remains partial. The second issue has to do with the specific discourse relation senses that polysemous discourse connectives may be used to express in context. This issue has not yet been systematically addressed in existing cohesion research. In TAACO 2.0, for example, the list of temporal connectives includes all occurrences of *while*, even though it does not always mark a temporal relation. A more systematic approach to addressing these two ambiguity issues will help improve and validate the reliability of connective-based cohesion indices.

### Current study

Considering the research gaps discussed previously, this study proposes a comprehensive set of 34 sense-aware connective-based cohesion indices and evaluates their correlations with and predictive power for cohesion ratings of ELLs’ written production in comparison to 25 existing connective-based indices. Within the context of the current study, we use the term “discourse relation sense” to refer to the specific type of discourse relation signaled by a discourse connective (see discussion in the Method section), and our sense-aware indices account for discourse versus non-discourse uses of connectives as well as the specific discourse relations expressed by explicit discourse connectives used in written texts. The two specific research questions addressed are:

1. How do existing and sense-aware connective-based cohesion indices correlate with cohesion ratings of young ELLs’ written production?
2. How do existing and sense-aware connective-based cohesion indices predict cohesion ratings of young ELLs’ written production?

### Method

#### ELL writing data

The writing data used in the current study consisted of the full training dataset of the Kaggle Feedback Prize English Language Learning Competition (Vanderbilt University & The Learning Agency Lab, 2022). The goal of the competition was to develop effective automated essay scoring and feedback tools for ELLs. The dataset was part of the larger English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus<sup>2</sup> released in 2023, which contains about 6,500 independent essays written on

<sup>1</sup><https://www.linguisticanalysisitools.org/taaco.html>

<sup>2</sup><https://github.com/scrosseye/ELLIPSE-Corpus>

44 different prompts by ELLs in the United States with diverse backgrounds in terms of gender, race/ethnicity, grade level, and economic disadvantage. The dataset used for the 2022 Kaggle competition and in the current study included 3,911 argumentative essays written by 8th to 12th grade ELLs in the United States as part of state standardized writing assessments from the 2018 to 2019 and 2019 to 2020 school years. This dataset contained a total of 1.683 million words, with an average of 430.372 words per essay (standard deviation [SD] = 191.974). Each essay was scored on a scale of 1.0 to 5.0 (with 0.5 increments) for each of the following six analytic measures: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The essays were rated by a pool of 26 raters recruited and trained by the corpus compilers. Most raters were senior undergraduate students or graduate students in an applied linguistics department, and all raters had experience teaching English as a second language. The corpus compilers adopted a double-blind rating process with 100% adjudication to ensure rating reliability, with each essay independently reviewed by two raters and adjudicated by a third one when necessary; a Many-Facet Rasch Measurement analysis was also conducted for the raters and texts to confirm the reliability of the ratings.<sup>3</sup> Only the cohesion scores were used in the current study. The rating rubric for all measures can be found at the ELLIPSE Corpus dataset,<sup>2</sup> which specifies that a cohesion score of 5 corresponds to the following: “Text organization consistently well controlled using a variety of effective linguistic features such as reference and transitional words and phrases to connect ideas across sentences and paragraphs; appropriate overlap of ideas.” The average cohesion scores of the essays in the dataset was 3.127 (SD = .663).

### *Connective identification and disambiguation*

Each text in the dataset was processed in three steps. First, we used LanguageTool,<sup>4</sup> an open-source grammar checker, to automatically correct spelling, grammar, and punctuation errors in each text. This step served to help minimize issues such errors might pose to syntactic parsing. Second, we used Stanford CoreNLP 3.6.0<sup>5</sup> (Manning et al., 2014) to perform constituency parsing on each text. This step outputs a constituency parse (i.e., a phrase-structure tree) for each sentence in the text that captures the hierarchical relations among the sentence’s constituents. Finally, we used the Explicit Discourse Connectives Tagger<sup>6</sup> (EDCT; Pitler & Nenkova, 2009) to automatically identify and disambiguate explicit connectives in each parsed text. The EDCT takes a constituency-parsed text as input and augments the constituency parses by annotating each connective with a tag indicating that it is either a non-discourse connective (Non-DC) or a discourse connective (DC) with a specific discourse relation sense.

The EDCT was trained and evaluated using data from the Penn Discourse Treebank (PDTB) Version 2.0 (Prasad et al., 2008), a version of the one-million-world Wall Street Journal Corpus annotated for discourse relations and their arguments. Prasad et al. (2008) understood discourse relations as holding “between two and only two arguments” and characterized arguments as “abstract objects” that are commonly expressed in single clauses or sentences but can also be associated with

<sup>3</sup><https://www.the-learning-agency-lab.com/the-feedback-prize-case-study/>

<sup>4</sup><https://pypi.org/project/language-tool-python/>

<sup>5</sup><https://stanfordnlp.github.io/CoreNLP/>

<sup>6</sup><https://www.cis.upenn.edu/~nlp/software/discourse.html>

multiple clauses or sentences or be denoted by non-clausal units such as discourse deictics (i.e., *this*, *that*) that refer to abstract objects or nominalizations with an event interpretation (e.g., *their failure to pass the exam*) (p. 2962). Discourse connectives are thus explicit linguistic expressions (both single words and multiple-word expressions) that signal discourse relations between two arguments. Each discourse relation was annotated by marking the discourse connective signaling it (e.g., *because*), labeling the discourse connective with its discourse relation sense, and annotating the attributes of its arguments. Altogether, the corpus contains annotations of 18,459 instances of 100 different explicit discourse connectives. The hierarchical taxonomy of discourse relation senses includes four top-level semantic classes, namely, Expansion (information in one clause elaborates that in the other), Contingency (information in one clause expresses the cause of that in the other), Comparison (information in one clause is compared or contrasted with that in the other), and Temporal (information in one clause is temporally related to that in the other) (see examples in Table 1). The PDTB also provides two lower-level (i.e., type and subtype) annotations for each top-level class. For example, Temporal has two types, namely, Asynchronous and Synchronous, and the latter has two subtypes, namely, Precedence and Succession. The full hierarchal taxonomy can be found in Prasad et al. (2008, p. 2965). The EDCT annotates each instance of a discourse connective with one of the four top-level discourse relation senses only. These categories are theoretically meaningful as they align with the four conjunctive relations (i.e., additive, causal, adversative, and temporal) distinguished in Halliday and Hasan's (1976) framework and concurred by Louwse (2002).

Pitler and Nenkova (2009) reported an accuracy of .9626 for distinguishing DCs versus Non-DCs and an accuracy of .9415 for classifying DCs into the four discourse relation senses. We evaluated the performance of the EDCT for tagging the essays produced by ELLs on 30 texts randomly sampled from our dataset, with 10 from each of the following three score bands: low (1–2 points, N = 352), mid (2.5–3.5 points, N = 2,874), and high (4–5 points, N = 685). Altogether, these 30 texts contained 1,005 connective tokens. One researcher and a trained graduate student assistant collaboratively annotated all 1,005 connective tokens manually. We first labeled each connective as either clear/proper (N = 899 or 89.5%) or unclear/improper (N = 106 or 10.5%). An instance was considered unclear/improper if the two annotators agreed that the discourse relation of the connective was either difficult to be determined from the context or improper for the context. For example, the annotators labeled the word *but* in “*We had this sorta thing to but easier*” as unclear/improper, which the EDCT labeled as Comparison. Notably, most unclear/improper instances were found in the 10 low-scoring samples, which represented 9.0% (i.e., 352 of 3,911) of the full dataset. In the 20 mid- and high-scoring samples, which represented 91.0% (i.e., 3,559 of 3,911) of the full dataset, fewer than 5% of the connectives were labeled as unclear/improper. Although in most cases the tags assigned to such unclear/improper instances by the EDCT appeared to align with the most likely senses intended by the learners, we decided to exclude them in reporting the accuracy of the EDCT to ensure the reliability of our evaluation. Next, we labeled each clear/proper instance with one of the five tags in the EDCT tagset (i.e., Non-discourse, Expansion, Contingency, Comparison, and Temporal). A comparison of the tags assigned by the EDCT and the annotators to the 899 clear/proper instances revealed an overall accuracy of 90.4% (i.e., 813 of 899) of the EDCT. More details about the accuracy of the EDCT on samples from each score band and its precision, recall, and F-score for each EDCT connective type can be found in Appendix S1.

**Table 1.** The 34 sense-aware connective-based cohesion indices proposed in the current study

Index name	Description	Example
Discourse connectives		
DC_token_density	density of discourse connective tokens	See examples for Expansion, Contingency, Comparison, and Temporal subsequently
DC_TTR	diversity of discourse connectives	
DC_type_density	density of discourse connective types	
DC_type_num	number of discourse connective types	
Non-discourse connectives		
Non-DC_token_density	density of non-discourse connective tokens	<i>There are some apples <b>and</b> oranges on the table.</i>
Non-DC_token_ratio	ratio of non-discourse connective tokens	
Non-DC_TTR	diversity of non-discourse connectives	
Non-DC_type_density	density of non-discourse connective types	
Non-DC_type_num	number of non-discourse connective types	
Non-DC_type_ratio	ratio of non-discourse connective types	
Discourse connectives: Comparison		
Comparison_token_density	density of comparison connective tokens	<i>I wanted to stay, <b>while</b> everyone else wanted to leave.</i>
Comparison_token_ratio	ratio of comparison connective tokens	
Comparison_TTR	diversity of comparison connectives	
Comparison_type_density	density of comparison connective types	
Comparison_type_num	number of comparison connective types	
Comparison_type_ratio	ratio of comparison connective types	
Discourse connectives: Contingency		
Contingency_token_density	density of contingency connective tokens	<i>I thought you were not interested <b>since</b> you never responded to the invitation.</i>
Contingency_token_ratio	ratio of contingency connective tokens	
Contingency_TTR	diversity of contingency connectives	
Contingency_type_density	density of contingency connective types	
Contingency_type_num	number of contingency connective types	
Contingency_type_ratio	ratio of contingency connective types	
Discourse connectives: Expansion		
Expansion_token_density	density of expansion connective tokens	<i>She went to the United States in 1960, <b>and</b> has lived there ever since.</i>
Expansion_token_ratio	ratio of expansion connective tokens	
Expansion_TTR	diversity of expansion connectives	
Expansion_type_density	density of expansion connective types	
Expansion_type_num	number of expansion connective types	
Expansion_type_ratio	ratio of expansion connective types	
Discourse connectives: Temporal		
Temporal_token_density	density of temporal connective tokens	<i>We have not seen each other <b>since</b> I visited him last fall.</i>
Temporal_token_ratio	ratio of temporal connective tokens	
Temporal_TTR	diversity of temporal connectives	
Temporal_type_density	density of temporal connective types	
Temporal_type_num	number of temporal connective types	
Temporal_type_ratio	ratio of temporal connective types	

### Sense-aware connective-based cohesion indices

Based on the annotations provided by the EDCT, we proposed 34 sense-aware connective-based indices of local cohesion. Different from the approaches taken in previous research, we distinguished DCs from Non-DCs and further differentiated four subcategories of DCs based on the discourse relation senses they were used to express in written texts. For Non-DCs and each of the four subcategories of DCs, we computed the following six indices for each text: (i) token density, the number of connective tokens of this category divided by the number of word tokens; (ii) TTR, a measure of connective



diversity calculated by dividing the number of connective types of this category by the number of connective tokens of this category; (iii) type density, the number of connective types of this category divided by the number of word types; (iv) type number, the number of connective types of this category; (v) token ratio, the ratio of connective tokens of this category among all connective tokens; and (vi) type ratio, the ratio of connective types of this category among all connective types. For DCs, we computed the first four indices only, as the token and type ratios would be redundant of those for Non-DCs. Table 1 summarizes the 34 proposed indices. The 34 indices were computed for each text using a script written in Python 3. The Python scripts as well as all the extracted indices are openly available at Github.<sup>7</sup>

### *Connective-based cohesion indices from TAACO*

To evaluate the correlations with and predictive power for cohesion score of the 34 sense-aware connective-based cohesion indices proposed in the current study in comparison to and in combination with existing connective-based cohesion indices, we used all 25 connective-based cohesion indices from TAACO 2.0. According to Crossley et al. (2016b, pp. 1231–1232):

“Many of the connective indices are similar to those found in Coh-Metrix (McNamara et al., 2014) and are theoretically based on two dimensions. The first dimension contrasts positive versus negative connectives, and the second dimension is associated with the particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001), such as temporal, additive, and causative connectives.”

Table 2 summarizes the 25 connective-based indices from TAACO along with their descriptions and examples of the corresponding DCs involved. These indices are computed with a list-based approach. For a particular list, the frequency of occurrence of each list item in the text is counted, and the sum of the frequencies of all list items are then tallied and divided by the total number of words in the text, with the exception that some words are disambiguated using the Stanford dependency parser (Chen & Manning, 2014). More detailed descriptions of these indices can be found in Crossley et al. (2016b, 2019) and in the TAACO 2.0 manual.

### *Statistical analysis*

To address research question 1, we performed Pearson correlation analyses between the 25 connective-based indices from TAACO and cohesion score as well as between the 34 sense-aware connective-based indices proposed in the current study and cohesion score. In addition to statistical significance ( $p < .05$ ), we interpret correlation coefficients with at least a small effect size ( $|r| \geq .1$ ) as meaningful (Cohen, 1988). To address research question 2, we performed three sets of regression analyses. In the first set of analysis, all connective-based indices from TAACO were used to predict cohesion score. In the second set of analysis, all sense-aware connective-based indices were used to predict cohesion score. In the third set of analysis, all connective-based indices from TAACO and all sense-aware connective-based indices were used to predict cohesion score.

<sup>7</sup><https://github.com/iris2hu/sense-aware-cohesion/>

**Table 2.** The 25 connective-based cohesion indices from TAACO

Index name	Description	Examples
basic_connectives	number of basic connectives	<i>for, and, nor</i>
conjunctions	number of conjunctions	<i>and, but</i>
disjunctions	number of disjunctions	<i>or</i>
lexical_subordinators	number of lexical items functioning as subordinators	<i>after, although, as</i>
coordinating_conjuncts	number of coordinating conjuncts	<i>yet, so, nor</i>
addition	number of addition words	<i>and, also, besides</i>
sentence_linking	number of sentence linking words	<i>nonetheless, therefore, although</i>
order	number of order words	<i>to begin with, next, first</i>
reason_and_purpose	number of reason and purpose words	<i>therefore, that is why, for this reason</i>
all_causal	number of causal connectives	<i>although, arise, arises</i>
positive_causal	number of positive causal connectives	<i>arise, because, enabling</i>
opposition	number of opposition words	<i>but, however, nevertheless</i>
determiners	number of determiners	<i>a, an, the</i>
all_demonstratives	number of demonstratives	<i>this, that, these</i>
attended_demonstratives	number of demonstratives followed by a noun phrase	<i>THIS SENTENCE is an example</i>
unattended_demonstratives	number of demonstratives functioning as a noun phrase	<i>THIS is an example</i>
all_additive	number of additive connectives	<i>after all, again, all in all</i>
all_logical	number of logical connectives	<i>actually, admittedly, after all</i>
positive_logical	number of positive logical connectives	<i>actually, after all, all in all</i>
negative_logical	number of negative logical connectives	<i>admittedly, alternatively, although</i>
all_temporal	number of temporal connectives	<i>a consequence of, after, again</i>
positive_intentional	number of positive intentional connectives	<i>by, desire, desired</i>
all_positive	number of positive connectives	<i>actually, after, again</i>
all_negative	number of negative connectives	<i>admittedly, alternatively, although</i>
all_connective	number of all connectives	<i>actually, admittedly, after</i>

As noted by one reviewer, some scholars have pointed out that the commonly used procedures of predictor preselection based on bivariate correlations and of stepwise variable selection in regression modeling could result in the exclusion of useful predictors that may affect the outcome variable together with other predictors (e.g., Smith, 2018; Sun et al., 1996). Ferenci (2017) recommended that all potential confounders be included, or their selection be blinded to the outcome when constructing models. In light of these concerns and recommendations, in each of the three sets of analysis, we report the results from four regression models that do not require feature preselection and that collectively capture both linear and nonlinear relationships between the predictors and the outcome variable. These include the linear regression model without feature preselection, two Bayesian regression models that address multicollinearity and overfitting through regularization (i.e., Bayesian automatic relevance determination regression and Bayesian ridge regression), and a nonlinear ensemble-learning model (i.e., random forest regression). These models were built using the *scikit-learn* library in Python 3 ([https://scikit-learn.org/stable/user\\_guide.html](https://scikit-learn.org/stable/user_guide.html)). In our experiments, we

randomly divided the dataset into a training set (2,620 essays, or two thirds) and a test set (1,291 essays, or one third). Each model was trained on the training set and used to predict the cohesion scores of the essays in the test set. The performance of these models on the test set was evaluated using the Pearson correlation coefficient and root mean squared error (RMSE) between the predicted and actual cohesion scores as well as  $R^2$  and adjusted  $R^2$ .

## Results

### Correlation analysis

Table 3 presents the descriptive statistics of the 25 connective-based indices from TAACO as well as their correlations with cohesion scores, ranked by the absolute values of the correlation coefficients. The distribution plots of these indices are provided in Appendix S2. Among the 25 indices, only three exhibited significant and meaningful correlations ( $|r| \geq .1$ ,  $p < .05$ ) with cohesion scores, namely, number of sentence linking words ( $r = -.123$ ,  $p < .001$ ), number of positive causal connectives ( $r = -.110$ ,  $p < .001$ ), and number of basic connectives ( $r = -.101$ ,  $p < .001$ ). These correlations were all negative and small. Another 10 indices exhibited significant but trivial correlations ( $|r| < .1$ ,  $p < .05$ ).

Table 4 presents the descriptive statistics of the 34 sense-aware connective-based indices proposed in the current study as well as their correlations with cohesion scores,

**Table 3.** Descriptive statistics of the 25 connective-based indices from TAACO and their correlations with cohesion scores

Num	Index name	Mean	SD	Skewness	Kurtosis	<i>r</i>	<i>p</i>
1	sentence_linking	.038	.014	.489	.678	<b>-.123</b>	< .001
2	positive_causal	.033	.015	.895	2.129	<b>-.110</b>	< .001
3	basic_connectives	.044	.016	.559	1.548	<b>-.101</b>	< .001
4	all_logical	.057	.018	.574	.733	-.090	< .001
5	all_causal	.022	.012	1.011	2.156	-.077	< .001
6	disjunctions	.007	.007	1.445	2.814	-.074	< .001
7	conjunctions	.033	.014	.661	1.721	-.066	< .001
8	order	.006	.008	2.795	12.696	.066	< .001
9	all_negative	.014	.009	.858	1.021	-.058	< .001
10	all_temporal	.012	.010	1.711	5.234	.052	.001
11	positive_logical	.031	.013	.670	.751	-.052	.001
12	reason_and_purpose	.016	.009	.787	.983	-.040	.012
13	all_connective	.084	.021	.253	.752	-.035	.031
14	opposition	.007	.006	1.253	3.264	-.029	.065
15	unattended_demonstratives	.019	.012	1.017	1.969	-.029	.067
16	all_demonstratives	.024	.014	.862	1.136	-.024	.139
17	determiners	.081	.027	.547	.724	-.021	.188
18	lexical_subordinators	.071	.020	.210	.621	-.014	.371
19	coordinating_conjuncts	.005	.005	1.766	5.569	.012	.469
20	negative_logical	.007	.006	1.128	2.114	-.009	.576
21	positive_intentional	.015	.011	1.342	2.681	-.009	.591
22	all_positive	.086	.022	.270	.656	-.005	.736
23	addition	.035	.014	.660	1.672	.002	.880
24	all_additive	.053	.016	.388	1.004	-.002	.916
25	attended_demonstratives	.005	.006	1.932	5.702	.001	.967

Note: Num = number. SD = standard deviation. Bolded *r* values indicate significant and meaningful correlations ( $|r| \geq .1$ ,  $p < .05$ ).

**Table 4.** Descriptive statistics of the 34 sense-aware connective-based indices proposed in this study and their correlations with cohesion scores

Num	Index name	Mean	SD	Skewness	Kurtosis	<i>r</i>	<i>p</i>
1	DC_type_num	6.831	3.090	.240	.291	<b>.381</b>	< .001
2	Expansion_type_num	2.257	1.392	.785	1.324	<b>.330</b>	< .001
3	Comparison_type_num	1.066	.926	.907	1.032	<b>.260</b>	< .001
4	Non-DC_type_ratio	.516	.159	.764	1.049	<b>-.234</b>	< .001
5	Contingency_type_num	2.268	1.165	.251	.208	<b>.222</b>	< .001
6	Temporal_type_num	1.239	1.090	.959	1.168	<b>.202</b>	< .001
7	Expansion_type_ratio	.160	.089	.395	.342	<b>.195</b>	< .001
8	Non-DC_type_density	.036	.012	.854	1.464	<b>-.180</b>	< .001
9	Non-DC_token_density	.056	.023	.835	1.227	<b>-.172</b>	< .001
10	DC_type_density	.035	.015	.316	.783	<b>.171</b>	< .001
11	Expansion_type_density	.012	.007	.933	1.760	<b>.167</b>	< .001
12	Comparison_type_ratio	.073	.061	.644	.384	<b>.164</b>	< .001
13	Non-DC_token_ratio	.582	.184	.194	-.111	<b>-.153</b>	< .001
14	Comparison_type_density	.005	.005	1.034	1.756	<b>.147</b>	< .001
15	Expansion_token_ratio	.141	.093	.802	1.542	<b>.141</b>	< .001
16	Comparison_TTR	.502	.406	.047	-1.559	<b>.133</b>	< .001
17	Temporal_type_ratio	.084	.069	.619	.239	<b>.120</b>	< .001
18	Comparison_token_ratio	.049	.049	.906	1.612	<b>.118</b>	< .001
19	Expansion_token_density	.013	.009	1.240	1.860	<b>.118</b>	< .001
20	DC_token_density	.039	.019	.213	.283	<b>.117</b>	< .001
21	Comparison_token_density	.005	.004	1.311	2.356	<b>.109</b>	< .001
22	Temporal_TTR	.489	.397	.104	-1.506	<b>.106</b>	< .001
23	Temporal_type_density	.006	.005	1.084	2.110	<b>.102</b>	< .001
24	Temporal_token_ratio	.062	.065	1.575	3.502	<b>.099</b>	< .001
25	Non-DC_type_num	7.217	3.095	.661	.510	<b>.090</b>	< .001
26	Non-DC_TTR	.359	.158	1.388	2.853	<b>-.087</b>	< .001
27	Temporal_token_density	.006	.006	1.738	4.434	<b>.078</b>	< .001
28	Expansion_TTR	.478	.294	.412	-.632	<b>.052</b>	.001
29	Contingency_token_ratio	.166	.108	.900	1.977	<b>.034</b>	.033
30	Contingency_type_ratio	.166	.085	.648	3.371	<b>.029</b>	.073
31	Contingency_type_density	.012	.007	.885	1.818	<b>.022</b>	.169
32	DC_TTR	.484	.211	.823	.910	<b>-.015</b>	.348
33	Contingency_token_density	.016	.010	.508	.395	<b>.015</b>	.356
34	Contingency_TTR	.446	.278	.699	-.239	<b>-.014</b>	.398

Note: Num = number. SD = standard deviation. Bolded *r* values indicate significant and meaningful correlations ( $|r| \geq .1$ ,  $p < .05$ ).

ranked by the absolute values of the correlation coefficients. The distribution plots of these indices are provided in [Appendix S2](#). Among these indices, 23 exhibited significant and meaningful correlations ( $|r| \geq .1$ ,  $p < .05$ ) with cohesion scores. Notably, two indices achieved medium effect sizes ( $|r| \geq .3$ ) (Cohen, 1988), namely, DC\_type\_num ( $r = .381$ ,  $p < .001$ ) and Expansion\_type\_num ( $r = .330$ ,  $p < .001$ ), and 16 indices exhibited stronger correlations than all 25 connective-based indices from TAACO (i.e.,  $|r| > .123$ ). Another six indices exhibited significant but trivial correlations ( $|r| < .1$ ,  $p < .05$ ).

### Regression analysis

Table 5 summarizes the performance on the test set of the four regression models trained on the training set with different sets of indices. With the 25 connective-based indices from TAACO, the cohesion scores predicted by the four regression models showed correlation coefficients ranging from .275 (Bayesian ridge regression)

**Table 5.** Performance of the four regression models on the test set

Model	Pearson's <i>r</i>	<i>p</i>	RMSE	<i>R</i> <sup>2</sup>	Adjusted <i>R</i> <sup>2</sup>
Models trained with the 25 connective-based indices from TAACO					
Linear	.285	<.001	.629	.079	.061
Bayesian ARD	.282	<.001	.629	.078	.060
Bayesian ridge	.275	<.001	.631	.074	.056
Random forest	.332	<.001	.620	.105	.087
Models trained with the 34 sense-aware connective-based indices					
Linear	.409	<.001	.599	.165	.142
Bayesian ARD	.411	<.001	.598	.168	.145
Bayesian ridge	.414	<.001	.597	.170	.148
Random forest	.402	<.001	.600	.161	.138
Models trained with all 59 connective-based indices					
Linear	.447	<.001	.588	.196	.158
Bayesian ARD	.445	<.001	.588	.196	.157
Bayesian ridge	.439	<.001	.589	.192	.153
Random forest	.436	<.001	.591	.188	.149

Note: ARD = automatic relevance determination. RMSE = root mean squared error.

to .332 (random forest regression) with human-rated cohesion scores. These were outperformed by all four regression models trained on the 34 sense-aware connective-based indices, whose predicted cohesion scores achieved correlation coefficients between .402 (random forest regression) and .414 (Bayesian ridge regression) with human-rated cohesion scores. When all 59 connective-based indices were included, the performance of the four regressions models further improved, with their predicted cohesion scores showing correlation coefficients ranging from .436 (random forest regression) to .447 (linear regression) with human-rated cohesion scores. The same patterns of performance changes were observed for RMSE, *R*<sup>2</sup>, and adjusted *R*<sup>2</sup>. Overall, the models trained on the sense-aware connective-based indices performed better in predicting the cohesion scores on the test set than those trained on the connective-based indices from TAACO, and the models trained on all 59 indices achieved yielded the most accurate predictions.

## Discussion

In light of the observation that existing cohesion indices have not yet systematically addressed issues related to lexical ambiguity, this study proposed a set of 34 sense-aware connective-based indices of cohesion based on the annotations provided by the Explicit Discourse Connective Tagger (Pitler & Nenkova, 2009), which allowed us to differentiate discourse versus non-discourse uses of explicit connectives as well as the specific discourse relation senses expressed by discourse connectives in context. We further examined their correlations with and predictive power for cohesion ratings of argumentative essays written by 8th to 12th grade ELLs in the United States both in comparison to and in combination with 25 connective-based indices from TAACO 2.0. Our analyses yielded a number of substantive findings with useful implications for cohesion research.

The results pertaining to our two research questions indicate that the sense-aware connective-based indices of cohesion are more strongly correlated with and better predictors for cohesion scores than existing connective-based indices. Three of the 25 connective-based indices from TAACO exhibited significant and meaningful

correlations with cohesion scores, all with small effect sizes. In contrast, 23 of the 34 sense-aware connective-based indices proposed in the current study showed significant and meaningful correlations with cohesion scores, two with medium effect sizes and 21 with small effect sizes. Among these, 16 sense-aware indices showed stronger correlations with cohesion scores than all 25 TAACO indices. The four regression models trained with the 34 sense-aware connective-based indices on the training set all performed better in predicting the cohesion scores on the test set than all four regression models trained with the 25 connective-based indices from TAACO.

The stronger correlational relationship with and greater predictive power for cohesion score achieved by the sense-aware connective-based indices may be attributed to two main factors. First, the TAACO indices were all based on normalized frequency counts (similar to our density indices), whereas the sense-aware indices captured several additional aspects of connective use, including total type frequency counts, connective diversity (i.e., the TTR indices), and connective complexity (i.e., the ratio indices). Our findings showed that these different aspects of connective use all contributed useful information for assessing cohesion. As noted by one reviewer, the indices based on total type frequency counts are likely affected by text length and may therefore be measuring both text length and cohesion. These indices were examined in the current study along with the density indices because they could more directly reflect the full range of discourse connectives produced by learners than the density indices. For example, a 100-word text with 10 connective types and a 150-word text with 15 connective types would have the same *DC\_type\_density* (i.e., 10 per 100 words). Although normalization accounts for text length, it may also conceal some differences in the learners' productive ability, and it is an empirical question whether raters may be sensitive to the actual range of connective types produced by the learners writing for the same tasks. The weak correlation between text length (i.e., number of words per text) and cohesion scores ( $r = .222, p < .001$ ) suggests that some of the high correlations exhibited by some of *type\_num* indices could not all be the result of increased length but also reflected the raters' sensitivity to the absolute range of connective types used. Lu (2012) also found that when evaluating timed speech samples produced by L2 speakers on the same tasks, the number of different words, a measure of the full range of word types produced, usefully complemented lexical diversity indices based on normalized frequencies in predicting quality ratings. Critically, however, the use of such *type\_num* indices should only be considered along with density indices when evaluating samples produced for the same writing task(s) using the same rubric, as Lu (2012) also recommended.

Second, and more importantly, the discourse relation sense disambiguation of the connectives likely helped improve the reliability of the connective-based indices. A review of the lists of different types of connectives used to compute the connective-based indices in TAACO shows that the lexical ambiguity of connective word forms could introduce noise into those indices. With only partial ambiguity resolution of some polysemous connective word forms, a connective word form with a non-discourse use may be counted as one with a discourse use, and a discourse connective used with one discourse relation sense may be counted as one used with a different discourse relation sense. Such noise could affect the reliability of the indices computed and subsequently weaken their ability to assess what they were designed to measure, namely, local cohesion achieved through the use of connectives. Examples 1 and 2 illustrate how the EDCT helped us address these issues by tagging Non-DCs with the #Non-DC tag and differentiating among different discourse relation senses of DCs. For instance, In Example 1, the EDCT tagged the first instance of *and* as a Non-DC and

the second instance as an Expansion connective. In Example 2, it tagged the first instance of *while* as a Temporal connective (and, importantly, not a Comparison connective) and the second instance as a Non-DC.

**Example 1.** *The school board shouldn't extend the school day by adding one and#Non-DC a half hours because#Contingency the students won't have time for#Non-DC themselves at home, the classes would take longer, and#Expansion the school will end at evening time.*

**Example 2.** *For example#Expansion, sometimes even I get distracted while#Temporal doing my work at home... If#Contingency you are at home for#Non-DC 8 hours doing work, you wouldn't want to spend time relaxing at your home because#Contingency eventually it would get boring there after#Non-DC a while#Non-DC.*

A comparison of the TAACO indices and the sense-aware density indices, both based on normalized frequency counts, can shed some light on the positive effect of discourse relation sense disambiguation on the connective-based indices. As indicated in Tables 3 and 4, nine of the 12 sense-aware density indices but only three of the 25 TAACO indices achieved significant and meaningful correlations ( $|r| \geq .1$ ,  $p < .05$ ) with cohesion scores. Furthermore, five sense-aware density indices showed stronger correlations than all 25 TAACO indices. These differences suggest a positive effect of discourse relation sense disambiguation on the connective-based indices. To further isolate the effect of discourse relation sense disambiguation, we calculated the non-sense-aware counterparts of the following four sense-aware indices reflecting overall discourse connective use: number of DC types, density of DC types and tokens, and TTR. To this end, we generated a list of all connectives based on the tags assigned by the EDCT and used this list to obtain frequency counts of connective types and tokens in each sample, similar to how most connectives were counted in TAACO. This means Non-DCs were counted as discourse connectives as well. Table 6 presents the descriptive statistics of the non-sense-aware and sense-aware indices of overall connective use and their correlations with cohesion scores. Without differentiating Non-DCs from DCs, the three non-sense-aware frequency and density indices all exhibited higher means and SDs than their sense-aware counterparts. The lower non-sense-aware TTR value could be attributed to the larger effect of overcounting on connective tokens than on connective types. Only one non-sense-aware index (vs. three sense-aware indices) achieved significant and meaningful correlations with cohesion scores ( $|r| \geq .1$ ,  $p < .05$ ). The three sense-aware frequency and density indices all showed stronger correlations with the cohesion scores than their non-sense-aware counterparts. These findings further confirm the positive effect of discourse relation sense disambiguation on

**Table 6.** Descriptive statistics of non-sense-aware and sense-aware indices of overall connective use and their correlations with cohesion scores

Index	Non-sense-aware		Sense-aware	
	Mean (SD)	<i>r</i>	Mean (SD)	<i>r</i>
DC_token_density	.104 (.029)	-.038*	.039 (.019)	.117***
DC_ttr	.295 (.110)	-.010	.484 (.211)	-.015
DC_type_density	.079 (.022)	.028	.035 (.015)	.171***
DC_type_num	10.264 (3.441)	.359***	6.831 (3.090)	.381***

Note: DC = discourse connective. SD = standard deviation. \*  $p < .05$ , \*\*\*  $p < .001$ .

connective-based indices. They also echo those from Lu and Hu (2022), who reported superior performance of sense-aware frequency-based indices of lexical sophistication that accounted for the specific senses with which polysemous words are used for predicting holistic ratings of L2 English writing quality over existing frequency-based lexical sophistication indices that did not account for lexical ambiguity.

The three TAACO indices that were significantly and meaningfully correlated with cohesion scores, namely, sentence linking connectives (e.g., *although, therefore, for, then, while, so, since, as, after*), positive causal connectives (e.g., *arise, cause, condition, consequence, make, result*), and basic connectives (i.e., *for, and, nor, but, or, yet, so*), all exhibited negative correlations. This result suggests that argumentative essays with higher cohesion ratings would contain fewer such connectives. Multiple previous studies have also reported negative correlations between connective-based indices of local cohesion and cohesion, coherence, or quality ratings of L1 and L2 writing and have often explained the negative correlations with the increased use of other types of local cohesive devices (e.g., lexical/semantic overlap and semantic similarity) in higher-rated writing or by more advanced learners in place of connectives (Crossley & McNamara, 2010, 2011, 2012; Guo et al., 2013; Kim & Crossley, 2018). Meanwhile, among the 23 sense-aware connective-based indices that showed significant and meaningful correlations with cohesion scores, only four Non-DC indices (i.e., the ratio and density of Non-DC types and tokens) exhibited negative correlations, whereas the other 19 DC-based indices all exhibited positive correlations. These results suggest that the negative correlations reported between connective-based indices and cohesion or coherence ratings in previous studies could have arisen from the noise in those indices with the confusion of discourse and non-discourse uses of connective word forms and the different discourse relation senses of discourse connectives. When these confusions were removed, higher-rated argumentative essays produced by young ELLs tended to use more instead of fewer DCs, as indicated by the positive correlations between cohesion scores and the three indices gauging the number of DC types and the density of DC types and tokens. The TTR of DCs, however, showed no significant correlation with cohesion scores. In terms of the four subtypes of DCs, indices related to Comparison and Expansion DCs exhibited stronger correlations with cohesion scores than those related to the other two types, with positive correlations found for all six indices related to the former and five indices related to the latter (i.e., all but the TTR of Expansion DCs). These findings show that higher-rated essays contained more Comparison and Expansion DC types, greater ratios and density of Comparison and Expansion DC types and tokens, and more diverse Comparison DCs. Furthermore, four indices related to Temporal DCs also showed significant correlations (i.e., all but the ratio and density of Temporal DC tokens), indicating that higher-rated essays also contained a greater number, ratio, and density of Temporal DC types as well as more diverse Temporal DCs. Finally, indices related to Contingency DCs showed the weakest correlations with cohesion scores overall, with positive correlations found only for the number of Contingency DC types, whereas the other five showed no meaningful correlations. All in all, these findings show that increased uses of the four subtypes of DCs did not negatively affect cohesion score in any way but positively affected cohesion score in several ways in our dataset.

The models trained on all TAACO and sense-aware indices outperformed those trained on either set of indices alone, indicating that the two sets of indices can complement each other in useful ways. In particular, the subcategories of connectives involved in the three TAACO indices with significant and meaningful correlations with cohesion scores are either different from (sentence linking connectives and basic



connectives) or more fine-grained (positive causal connectives) than the subcategories differentiated in the sense-aware indices. These findings suggest that the sense-aware indices could be potentially enhanced with additional and/or more fine-grained ways for categorizing the discourse connectives.

In summary, our analysis has provided evidence for the value of addressing discourse relation sense ambiguity and integrating discourse relation sense information in connective-based indices of cohesion. The most important implication of the current study for future cohesion research is to systematically distinguish discourse and non-discourse uses of connective word forms and the specific discourse relation senses with which DCs are used in context. By extension, other types of cohesion indices, such as lexical/semantic overlap indices of local and global cohesion, may benefit from word sense disambiguation as well, as the overlap or repetition of two identical or potentially synonymous words with two different or unrelated senses does not contribute to cohesion. Theoretically, our findings on the predictive power of models trained with sense-aware connective-based indices for cohesion score add evidence to prior SLA literature on the role of local cohesion in L2 written production. Along with similar findings from previous studies, the differential degrees of correlations between the indices based on different types of connectives and cohesion score offer empirical support for the usefulness of taxonomies of discourse or conjunctive relations in analyzing local cohesion of L2 writing (Halliday & Hasan, 1976; Louwerse, 2002; Prasad et al., 2008). Meanwhile, given the discrepancy in the polarity of the correlational relationship with cohesion scores found for existing and sense-aware connective-based cohesion indices, it would appear useful to revalidate some of the findings reported in previous research on the correlational relationship between connective-based indices of local cohesion and cohesion, coherence, or quality ratings of adult L1 and L2 writing. Our findings also confirm the importance of appropriate use of DCs expressing different discourse relations in writing pedagogy and assessment for young ELLs. In particular, the types and ratios of Expansion and Comparative connectives exhibited stronger correlations with cohesion score than those of Contingency and Temporal connectives, suggesting potentially greater relative importance of Expansion and Comparative connectives in achieving local cohesion in argumentative writing. It is also important to help learners to expand the repertoire of connectives of each type and to develop the ability to choose appropriate and diverse connectives to express precise discourse relations in context without relying on repetitive uses of a small set of basic ones.

## Conclusion

This study proposed 34 new sense-aware connective-based cohesion indices that considered the discourse and non-discourse uses of connective word forms and the specific discourse relation senses of discourse connectives in context. To this end, we used the Explicit Discourse Connective Tagger to identify explicit connectives from the text and tag each explicit connective as either a Non-DC or a DC with one of four discourse relation senses (i.e., Comparison, Contingency, Expansion, Temporal). This allowed us to distinguish DCs more systematically from Non-DCs and determine the specific discourse relation sense expressed by each DC in context. Results showed that 16 of the 34 indices we proposed exhibited stronger correlations with cohesion ratings of argumentative essays produced by young ELLs than all 25 connective-based indices from TAACO, that models trained with the sense-aware indices showed stronger

predictive power for cohesion score than those trained with the connective-based indices from TAACO, and that models trained with all TAACO and sense-aware indices exhibited greater predictive power for cohesion score than those trained using either set of indices alone. Our findings highlight the value of integrating sense-level information in developing cohesion indices. As one reviewer suggested, future research could train or fine-tune large language models using the PDTB to improve the accuracy and granularity (using the full hierarchical taxonomy adopted in the corpus) of discourse relation sense tagging. Future cohesion research could also apply sense-aware indices to re-examine the relationship of connective-based cohesion indices to cohesion, coherence, or quality ratings of writing produced by other learner populations and/or for other types of writing tasks and, more importantly, investigate how sense disambiguation could be integrated into other types of cohesion indices based on word forms, such as lexical and semantic overlap indices of local and global cohesion.

**Supplementary material.** The supplementary material for this article can be found at <http://doi.org/10.1017/S0272263124000202>.

**Acknowledgments.** This research was supported by a grant from the Center for Language Education and Cooperation at the Ministry of Education of China (No. 22YH04ZW), a grant from the National Language Commission of China (No. ZDA145-9), and a grant from the Beijing Federation of Social Science Circles (No. 21DTR037). It is also supported by the Fundamental Research Funds for the Central Universities of China.

**Competing interest.** The authors declare no competing interests.

**Data availability statement.** The experiment in this article earned Open Data and Materials badges for transparent practices. The data and materials are available at <https://github.com/iris2hu/sense-aware-cohesion>.

## References

- Biler, A. (2018). *The role of cohesion in second language reading comprehension* (Unpublished doctoral dissertation). University of South Carolina, Columbia, SC.
- Bilki, Z. (2014). *A close observation of second language (L2) readers and texts: meaning representation and construction through cohesion* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.
- Bayraktar, H. (2011). *The role of lexical cohesion in L2 reading comprehension: Awareness of lexical cohesive links and L2 reading test performance*. VDM Verlag Dr. Müller.
- Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. <https://doi.org/10.3115/v1/d14-1082>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236–1241). Cognitive Science Society.
- Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51, 14–27. <https://doi.org/10.3758/s13428-018-1142-4>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16. <https://doi.org/10.1016/j.jslw.2016.01.003>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48, 1227–1237. <https://doi.org/10.3758/s13428-015-0651-7>

- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, 35, 115–135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>
- Degand, L., & Sanders, T. (2002). The impact of relational markers on expository text comprehension both in L1 and L2. *Reading and Writing*, 15, 739–757. <https://doi.org/10.1023/A:1020932715838>
- Ferenci, T. (2017). Variable selection should be blinded to the outcome. *International Journal of Epidemiology*, 46, 1077–1079. <https://doi.org/10.1093/ije/dyx048>
- Graesser, A. C., McNamara, D. S., Louwse, M. M., & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36, 193–202. <https://doi.org/10.3758/bf03195564>
- Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18, 218–238. <https://doi.org/10.1016/j.asw.2013.05.002>
- Halliday, M., & Hasan, R. (1976). *Cohesion in English*. Longman. <http://archives.unc.edu.dz/handle/123456789/111528>
- Jonz, J. (1987). Textual cohesion and second language comprehension. *Language Learning*, 37, 409–438. <https://doi.org/10.1111/j.1467-1770.1987.tb00578.x>
- Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing*, 37, 39–56. <https://doi.org/10.1016/j.asw.2018.03.002>
- Louwse, M. M. (2001). *From cohesion in text to coherence in comprehension* (Doctoral dissertation). University of Edinburgh. <https://era.ed.ac.uk/handle/1842/22424>
- Louwse, M. M. (2002). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 12, 291–315. <https://doi.org/10.1515/cogl.2002.005>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96, 190–208. [https://doi.org/10.1111/j.1540-4781.2011.01232\\_1.x](https://doi.org/10.1111/j.1540-4781.2011.01232_1.x)
- Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods*, 54, 1444–1460. <https://doi.org/10.3758/s13428-021-01675-6>
- Manning, C. D., Surdeanu, M., Bauer, J. A., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, MD. <https://doi.org/10.3115/v1/p14-5010>
- McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Matrix*. Cambridge University Press.
- Pitler, E., & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore. <https://doi.org/10.3115/1667583.1667589>
- Prasad, R. B., Dinesh, N., Lee, A. T. K., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (pp. 2961–2968). <https://people.cs.pitt.edu/~huynv/research/argument-mining/The%20Penn%20Discourse%20TreeBank%20.pdf>
- Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5, Article number 32. <https://doi.org/10.1186/s40537-018-0143-6>
- Sun, G., Shook, T., & Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, 49, 907–916. [https://doi.org/10.1016/0895-4356\(96\)00025-x](https://doi.org/10.1016/0895-4356(96)00025-x)
- Vanderbilt University, & The Learning Agency Lab. (2022). *Kaggle Feedback Prize English Language Learning Competition* [Dataset]. Retrieved from <https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data>
- Zhang, X., Lu, X., & Li, W. (2022). Beyond differences: Assessing effects of shared linguistic features on L2 writing quality of two genres. *Applied Linguistics*, 43, 168–195. <https://doi.org/10.1093/applin/amab007>

---

**Cite this article:** Lu, X., & Hu, R. (2024). Sense-aware connective-based indices of cohesion and their relationship to cohesion ratings of English language learners' written production. *Studies in Second Language Acquisition*, 46: 644–662. <https://doi.org/10.1017/S0272263124000202>